TEST VALIDITY

As described in the AERA, APA, and NCME *Standards for Educational and Psychological Testing* (1999), "Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests.... The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations" (p. 9). Various types of evidence may be considered in establishing the validity of test scores, and a number of methods are typically used to gather such evidence.

The validation process used by Evaluation Systems group of Pearson followed professionally accepted procedures for the validation of licensure/certification tests. The validation process focused primarily on establishing that the content of the tests was appropriate for the purpose of the testing program. In addition, Evaluation Systems provided guidance to test takers, teacher preparation programs, and statewide stakeholders regarding the appropriate interpretation and use of program test scores.

Throughout the various steps of test development, Evaluation Systems aimed to enhance the validity of the tests as recommended by the *Standards for Educational and Psychological Testing* (1999). Steps taken by Evaluation Systems included

- **Establishing the basis for the test.** The purpose of the testing program—to support state educator licensure decisions—and the test areas to be assessed were established by state rules and regulations.
- **Defining the test objectives.** AERA, APA, and NCME Standard 14.4 states that "evidence of validity based on test content requires a thorough and explicit definition of the content domain of interest" (p. 160). The test objectives described the content knowledge that the practitioner must possess to practice appropriately and, therefore, defined eligible test content. These test objectives were reviewed, revised, and approved by practicing educators and faculty at educator preparation institutions.
- **Conducting content validation of the test objectives.** The "validation of credentialing tests depends mainly on content-related evidence, often in the form of judgments that the tests adequately represent the content domain of the occupation" (AERA, APA, & NCME, 1999, p. 157). Content validation of the test objectives occurred [not needed with "as well as" construction] through correlation with documentation of content requirements as well as through a survey of job incumbents.

    1. Test objectives were aligned with relevant laws and regulations and student and national standards, where available, to provide documentation of the basis of the test objectives. Thus, the content of the tests was verified as being relevant.
    2. A Content Validation Survey of the proposed test objectives was conducted among public school educators and college and university faculty. The survey asked educators to make judgments for each proposed test objective regarding its importance to the job of an educator in the state. High ratings of importance provided additional evidence regarding the validity of the content for the licensure assessment.

- **Validating test items.** The content of the test items on licensing tests should be determined by the requirements of the job(s) covered by the license. Test items were reviewed with specific reference to licensing and job requirements through reviews by the licensed practitioners and educators who served on the various advisory committees. During test item review meetings, committees of educators were asked to review each item and consider its alignment to the relevant validated test objective as well as its accuracy, freedom from bias, and job-relatedness.
- **Preventing bias.** The prevention of bias in a testing program is important as a matter of fairness and as an aspect of test validity. Guarding against bias in the test materials involved the collaboration of educators and reviewers focused on excluding language, content, or perspectives that might disadvantage examinees based on background characteristics irrelevant to the purpose of the test, and on including content and perspectives that reflect the diversity of a state's population. The Bias Review Committee (BRC) reviewed test materials for potential bias. In addition, educators from diverse backgrounds were invited to participate in the test development process. They served as members of Content Advisory Committees (CACs), reviewing the test objectives and draft test items for each test field. The Content Advisory Committees reviewed proposed test items and revised the content as necessary to ensure that the test items adequately covered the necessary subject matter knowledge and skills, and met the review criteria established for a state's testing program. Only those items that were accepted by these committees were considered for use on operational test forms.
- **Pilot testing of items.** Teacher licensure candidates participated in the pilot testing of questions proposed for the tests. Pilot test performance provided information that was used to gauge expected operational test performance. Acceptable item statistics based on pilot testing served as another source of evidence regarding the importance and relevance of the test content for educator licensure candidates.
- **Setting passing standards.** Another committee of educators was convened to help establish the passing standards for every test. These committees met to review the test content and provide recommendations of the level of performance deemed acceptable for entry-level educators. These judgments were then presented to the licensing agency for consideration in establishing passing scores at a level appropriate to the profession and consistent with the mandate of the licensing agency to protect the health, safety, and welfare of the public.
- **Communicating appropriate interpretations with test users.** It is important that test scores are understood and used appropriately by the various potential users of the test results. Evaluation Systems includes an explanatory page of text with every examinee score report describing the included information. This information is also posted on the testing program website. Reports to educator preparation institutions include appropriate interpretive cautions. In addition, Evaluation Systems has worked closely with the state to provide guidance regarding the appropriate and psychometrically sound uses of the test scores.

TEST RELIABILITY

AERA, APA, and NCME (1999) define test reliability as "the consistency of measurements when the testing procedure is repeated" (p. 25). There are a number of statistics that may be used to estimate test reliability. In general, reported reliability values range from zero to one, with higher values indicating greater reliability of test scores. In a licensing context, reliability measures may be influenced by many factors, such as

- **Number of examinees.** In general, reliability estimates based on larger numbers of examinees are more stable than estimates based on smaller numbers. For this reason, reliability estimates are calculated for tests that are taken by one hundred or more examinees.
- **Test length.** Reliability estimates tend to be higher for tests with greater numbers of questions.
- **Test content.** Reliability estimates are typically higher for tests that cover narrow, homogeneous content than for tests (such as many used for educator licensure) that cover a broad range of content.
- **Examinees' knowledge.** Reliability estimates tend to be higher if examinees in the group have widely varying levels of knowledge and lower if they tend to have similar levels of knowledge.

Total Test Decision Consistency. In a licensing context, the most important testing outcome is the pass/fail decision. Total test decision consistency is a reliability statistic that describes the consistency of the pass/fail decision on the total test. A single-test estimate of total test decision consistency (Breyer and Lewis, 1994) is provided for test forms taken by 100 or more examinees. Each test form is carefully divided to create two halves that are parallel in terms of item content and item statistics. Performance on the two test halves is then compared to provide a decision consistency statistic. This statistic is reported in the range of 0.00 to 1.00; the closer the estimate is to 1.00, the more consistent (reliable) the decision is considered to be.

Assessments with component subtests. This program includes assessments that consist of two or more subtests. The subtest model is used for two reasons. First, many educator licenses require candidates to demonstrate proficiency across a variety of domain content. With a single test model in which the total test score is based on performance on all test items, outstanding performance on one component (e.g., reading/language arts) may compensate for poorer performance on another (e.g., mathematics). In a subtest model, candidates must pass each subtest separately, thus providing evidence of acceptable proficiency on each component. Second, the subtest model allows candidates who pass some subtests and fail others to retake only the failed components. Both of these characteristics are considered advantages by many policy agencies. One consequence of the subtest model, however, is that the pass/fail decisions are based on a decreased number of test items, when compared to a total test model in which all test items contribute to the pass/fail decision. As a result, traditional reliability statistics tend to be much lower when computed at the subtest level, because reliability is a function of the number of test items. Thus, KR-20 and other reliability evidence cannot be expected to reach the levels found in tests of greater length.

Table 1

Illinois Certification Testing System (ICTS)

Test Field and Name	Number of Examinees*	Total Test Decision Consistency	Scorable Multiple-choice Items	Constructed-response Items	Weighting
055 English Language Proficiency	109	0.908	44	2	50/50
056 TLP - Spanish	863	0.914	44	2	50/50
101 APT: Birth to Grade 3	830	0.937	104	2	80/20
102 APT: Grades K-9	5342	0.933	104	2	80/20
103 APT: Grades 6-12	3264	0.922	104	2	80/20
104 APT: Grades K-12	5047	0.898	104	2	80/20
105 Science: Biology	445	0.894	100		
106 Science: Chemistry	293	0.868	100		
107 Early Childhood Education	1113	0.834	100		
109 Social Science: Economics	112	0.92	100		
110 Elementary/Middle Grades	6429	0.895	100		
111 English Language Arts	1119	0.928	100		
113 Social Science: Geography	162	0.912	100		
114 Social Science: History	1078	0.894	100		
115 Mathematics	928	0.899	100		
116 Science: Physics	171	0.859	100		
117 Social Science: Political Science	209	0.897	100		
118 Social Science: Psychology	208	0.907	100		
121 Social Science: Sociology and Anthropology	136	0.895	100		
135 Foreign Language: Spanish	407	0.901	80	2	67/33
142 Health Education	150	0.909	100		
143 Music	398	0.984	100		
144 Physical Education	1033	0.807	100		
145 Visual Arts	329	0.874	100		
155 Learning Behavior Specialist I	2668	0.931	100		
163 Special Education General Curriculum	2292	0.823	52		
171 Business, Marketing, and Computer Education	198	0.822	100		
174 Technology Education	107	0.817	100		
175 Library Information Specialist	179	0.973	100		
176 Reading Specialist	1169	0.942	100		
177 Reading Teacher	453	0.841	100		
301 Reading Comprehension	19562	0.811	38		
302 Language Arts	20316	0.814	34		
303 Mathematics	19229	0.841	28		
304 Writing	16194	1			

* 2010 - 2011 Program Year

Notes: Decision consistency indices are generated only for the test forms with at least 100 examinees during the program year. If more than one test form is included in the analysis for a field, reported values represent a weighted average across test forms. "Weighting" indicates the respective contributions of multiple-choice items and open-response items to the total test score.